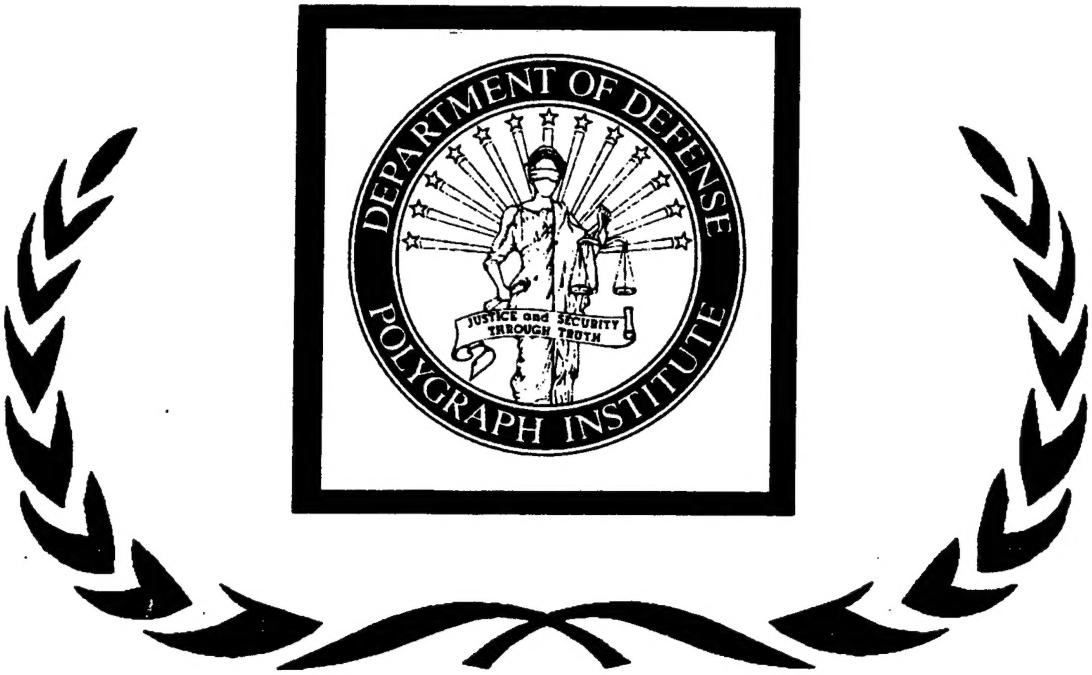


PolyScore 3.3 and Psychophysiological Detection of Deception Examiner Rates of Accuracy When Scoring Examinations from Actual Criminal Investigations

N. Joan Blackwell, M.S.

September 1998

Department of Defense Polygraph Institute
Fort McClellan, Alabama 36205-5114
Telephone: 205-848-3803
FAX: 205-848-5332

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1998	Final Report (Mar 96 - Sep 98)

4. TITLE AND SUBTITLE
PolyScore 3.3 and Psychophysiological Detection of Deception Examiner Rates of Accuracy when Scoring Examination from Actual Criminal Investigations

5. FUNDING NUMBERS
DoDPI96-P-0001

6. AUTHOR(S)
N. Joan Blackwell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Defense
Polygraph Institute
Building 3195
Fort McClellan, AL 36205-5114

8. PERFORMING ORGANIZATION REPORT NUMBER
DoDPI97-R-0006

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Department of Defense
Polygraph Institute
Building 3195
Fort McClellan, AL 36205-5114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
DoDPI97-R-0006
DoDPI96-P-0001

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Public release, distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A stratified random sample of 200 confirmed examinations were scored using PolyScore 3.3. Three experienced psychophysiological detection of deception (PDD) examiners scored the 100 Zone Comparison Test (ZCT) examinations, and three PDD examiners scored the 100 Modified General Question Test (MGQT) examinations, using a 7-position scale. The scores were converted to 3-position scale for comparison. PolyScore had an overall level of accuracy of 90.9% when scoring the ZCT examinations, but was less accurate when scoring the MGQT examinations. The PDD examiners had an overall level of accuracy of 82.3% and 73.3%--using the 7- and the 3-position scoring scales, respectively-when evaluating MGQT examinations, but were less accurate on ZCT examinations. A test for the significance of proportion differences was performed on the accuracy data and the differences between numerous comparisons were statistically significant. All computed Kappa values assessing the interrater agreement for the two groups of examiners were statistically significant. Finally, the proportion of concurrence between individual examiners, ranged from 76.8% to 81.0% for the ZCT examinations, and 78.8% to 92.0% for the MGQT examinations when using the 7-position scoring scale. Concurrence was lower when using the 3-position scoring scale.

14. SUBJECT TERMS	15. NUMBER OF PAGES
PolyScore, computerized scoring algorithms, 7-position scoring scale, 3-position scoring scale, psychophysiological detection of deception (PDD)	40
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	

Report No. DoDPI97-R-0006

PolyScore 3.3 and Psychophysiological Detection of Deception Examiner Rates of Accuracy When Scoring Examinations from Actual Criminal Investigations

N. Joan Blackwell, M.S.

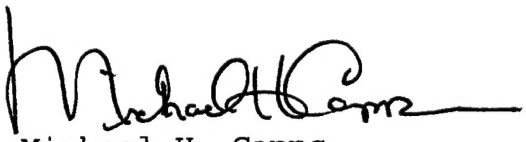
September 1998

Department of Defense Polygraph Institute
Fort McClellan, Alabama 36205

Director's Forward

The present psychophysiological detection of deception (PDD) study is a comparison between human raters and an automated analytical system blindly evaluating the same physiological data. Live criminal cases were used where ground truth was established independently. We are encouraged by the finding that the accuracy delivered by automated analysis compared favorably to that of the experienced human scorers. This is especially the case for true negatives (correct classification of truthful examinees), where the computer algorithm regularly outperformed the human scorers.

Computer algorithms have the advantage of perfect consistency, or reliability, while the outcomes from traditional manual scoring methods are influenced by the vagaries of individual differences. If the algorithm produces decisions that are as accurate as the average accuracy of multiple raters, it is to be preferred because of its consistency. Findings by DoDPI in previous studies, combined with related research performed elsewhere, suggest that validated automatic algorithms could play a greater role in some applications where blind scoring is now performed. DoDPI will continue to investigate and validate automated analytical systems to determine how they might best be employed.



Michael H. Capps
Director

Acknowledgments

The author wishes to express appreciation to the following organizations for providing manpower and resources in support of this research: the Air Force Office of Special Investigations, the Defense Investigative Service, the Department of Defense Polygraph Institute (DoDPI), the Naval Criminal Investigative Service, and the U. S. Army Criminal Investigations Division. Special thanks are extended to the examiners from each of those agencies who blind scored the examinations for this project, as well as, to the DoDPI staff members Ms. Joan Harrison-Woodard, Ms. Edna Knox, Ms. Charlene Stephens, and Ms. Rose Swinford for their roles in project coordination and/or data management.

This study was supported by funds from the DoDPI as project DoDPI96-P-0001. The views expressed in this report are those of the author and do not reflect the official policy or position of the Department of Defense or the U. S. Government.

Abstract

BLACKWELL, N. J. PolyScore 3.3 and Psychophysiological Detection of Deception Examiner Rates of Accuracy When Scoring Examinations from Actual Criminal Investigations. December 1997, Report No. DoDPI97-R-0006. Department of Defense Polygraph Institute, Fort McClellan, AL 36205.---A stratified random sample of 200 confirmed examinations were scored using PolyScore 3.3. Three experienced psychophysiological detection of deception (PDD) examiners scored the 100 Zone Comparison Test (ZCT) examinations, and three PDD examiners scored the 100 Modified General Question Test (MGQT) examinations, using a 7-position scale. The scores were converted to 3-position scale for comparison. PolyScore had an overall level of accuracy of 90.9% when scoring the ZCT examinations, but was less accurate when scoring the MGQT examinations. The PDD examiners had an overall level of accuracy of 82.3% and 73.3%--using the 7- and the 3-position scoring scales, respectively--when evaluating MGQT examinations, but were less accurate on ZCT examinations. A test for the significance of proportion differences was performed on the accuracy data and the differences between numerous comparisons were statistically significant. All computed Kappa values assessing the interrater agreement for the two groups of examiners were statistically significant. Finally, the proportion of concurrence between individual examiners, ranged from 76.8% to 81.0% for the ZCT examinations, and 78.8% to 92.0% for the MGQT examinations when using the 7-position scoring scale. Concurrence was lower when using the 3-position scoring scale.

Key-words: PolyScore, computerized scoring algorithms, 7-position scoring scale, 3-position scoring scale, psychophysiological detection of deception (PDD).

Table of Contents

List of Figures

List of Tables

For more than fifty years, the data resulting from modern-day physiological detection of deception (PDD) examinations have relied upon the human interpretation of physiological data. As with any evaluation system which bases its decisions on the effectiveness of such a markedly subjective process, the PDD field has been troubled by an ever increasing number of critics from the scientific community--Furedy (1985, 1987), Iacono (1991), Lykken (1981, 1986, 1988, 1991), and Raskin (1979, 1988) --to name just a few. The advent of computerized polygraph systems within the last decade, however, may have given rise to a new class of interpretive process, which could help to quiet the debate regarding the credible and dependable scoring of the PDD examination. One such interpretive process is the computerized scoring algorithm. PolyScore 3.3 is an example of such an algorithm. This study was conducted in order to assess and compare the level of accuracy generated by the PolyScore 3.3 algorithm and a representative group of certified PDD examiners.

PolyScore 3.3 is a user-friendly, personal computer software package designed to eliminate subjectivity from the process of scoring and interpreting PDD examinations. The scoring algorithm, which is based on a logistic regression model, was developed by the Johns Hopkins University Applied Physics Laboratory (APL) under contract to the National Security Agency (NSA). The prototype system, known then as the Polygraph Automated Scoring System (PASS) - Version 2.0, was first made available for research purposes in early 1993, with the public release of PolyScore 2.3 coming less than a year later. In the intervening years, APL has continued to make refinements to the PolyScore algorithm, as evidenced in each subsequent release of the scoring software.

The scoring software works in conjunction with the Axciton Computerized Polygraph (Axciton Systems, Incorporated, Houston, TX), and the Lafayette Computerized Polygraph (Lafayette Instrument Company, Lafayette, IN). Both are stand-alone PDD systems which record the physiological data (i.e., respiration, electrodermal and cardiovascular) collected during a PDD examination. PolyScore, in turn, uses that physiological data to produce an overall "probability of deception" for the examination (Johns Hopkins University Applied Physics Lab, 1996).

Similarly, PDD examiners use the same channels of physiological data to generate a decision of deception indicated (DI), no deception indicated (NDI), or inconclusive (INC). Both the PDD examiners and PolyScore 3.3 use a set of rules, or interpretation guidelines, to arrive at a decision, however, the analysis methods used by PolyScore 3.3 differ from those used by PDD examiners (Johns Hopkins University Applied Physics Lab, 1996; DoDPI, 1995). Additionally, based on a combination of formal training, past experience, agency policy, etc., the methods used by PDD examiners also differ, to some extent, from

one examiner to another, which potentially accounts for some of the variability seen in blind scoring accuracy rates and interrater agreement analyses (Ansley, Garwood, & Barland, 1984; Barland, 1972; Ben-Shakhar, Lieblich, Bar-Hillel, 1982; Blackwell, 1994; Forman & McCauley, 1986).

A system capable of providing accurate decisions, while eliminating the subjectivity factor when scoring examinations offers obvious advantages to the PDD community. By its very nature, a computerized scoring algorithm adds reliability to the evaluation process. The same test, scored using the same criteria, will generate the same results, time and again. As mentioned earlier, a comparable statement cannot always be made when discussing a test scored by humans.

Though Kircher and Raskin (1988) found no significant differences between computer and human evaluations when scoring mock crime data, they did make a distinction in their findings by comparing a computer's capability to that of an "expert" human interpreter. Kircher, et al., (1988) did not specify what constitutes an "expert," however, it should be clear that while certainly qualified, not all field examiners currently conducting PDD examinations could be considered to be experts. Logic dictates that within a typical agency, at any given time, there will be working examiners exhibiting varying levels of training, expertise, case resolution experience--and subjectivity.

The PolyScore algorithm was developed and "trained" on polygraph examination data. Though a set of mock crime data was used as the test case early in the algorithm's development, APL researchers soon recognized that they were able to produce much better accuracy rates when using "live" data (i.e., data which had resulted from actual criminal investigations). Use of the field cases rather than the laboratory-generated mock crime data presented a distinct problem, however; the ground truth information, (i.e., whether the person being tested was guilty or innocent of the crime) necessary for accuracy assessments was not readily available in the field cases.

APL remedied that concern by establishing a two-component guideline for attributing ground truth during algorithm development: (1) use confirmed cases (i.e., cases which have been resolved via the confession of the examinee or someone else), and (2) include cases which have been assigned the same decision by the original examiner and two other experienced examiners appointed to blind score the tests (Capps, 1993). Thus, examinations judged either DI, NDI, or INC, were incorporated into the test case database. The developers initially defined the algorithm's level of accuracy (99.4%--INC decisions eliminated) as, PolyScore's rate of agreement with the decisions from both the resolved cases and the cases evaluated by the three examiners (Capps, 1993).

In an attempt to confirm the rates of accuracy achieved by APL, the Department of Defense Polygraph Institute (DoDPI) initiated a study using laboratory-generated mock crime data with known ground truth. Blackwell (1994) found an accuracy rate of only 79.0% (INC decisions eliminated) when scoring mock crime data with the prototype software, PASS 2.0. It should be noted that the PDD examiner accuracy rate was essentially the same (79.6%) as the algorithm's accuracy rate, and that the INC rate was 20.8% and 10.0%, respectively, for the algorithm and the PDD examiners.

There are inherent difficulties with generalizing the results of mock crime cases to those cases collected under field conditions. Due to an inability to create sufficient stress-inducing consequences for everyone involved in a controlled laboratory mock crime scenario, the occurrence of a certain percentage of false decisions is inevitable (Ansley, Garwood, & Barland, 1984). Therefore, neither the algorithm, nor the PDD examiners were expected to attain the level of accuracy normally associated with the performance of their respective tasks; however, the accuracy rates for both the algorithm and the PDD examiners was lower than anticipated.

A second DoDPI study examined the effects on accuracy caused by the various refinements to the PolyScore system. Using the data set from the original research project, the examinations were scored by four versions of the algorithm: 2.0, 2.3, 2.9, and 3.0. While many accuracy calculations did improve when scored with each subsequent version--to include surpassing the overall accuracy rate of the original examiner--the results were still not equivalent to the percentages obtained by APL when using "live" data (Blackwell, 1995).

Due to a unique situation associated with the development of the APL algorithm, there now exists a database of over 400 confirmed cases resulting from actual criminal investigations--all collected on the Axciton Polygraph System. A stratified random sample of those cases was selected for use in this study to assess both PolyScore 3.3 accuracy and the level of accuracy attained by a group of experienced PDD examiners performing what could be deemed as a quality control (QC) role, by blind scoring the same set of examinations.

Initially designed to score a specific type of PDD examination known as the Zone Comparison Test (ZCT), PolyScore has since been adapted to score other tests formats, such as the Modified General Question Test (MGQT). Both formats are categorized as control question tests (CQT), and have been in widespread use within the PDD field since their development in the 1960's (OTA, 1983; DoDPI, January 1994; DoDPI, November 1995). As a result, examinations using both the ZCT and the MGQT were selected for use in this study.

The primary intent for conducting the study was to use the resulting data to quantify the level of benefit to be gained by utilizing an automated scoring system such as PolyScore 3.3. The author hypothesized that there would be no statistically significant difference between the various comparisons of accuracy attained by PolyScore 3.3 and the PDD examiners (i.e., PolyScore 3.3 accuracy compared to examiner accuracy [innocent, guilty, and overall], and interrater agreement for both the ZCT and the MGQT test formats).

Method

During this study, three experienced PDD examiners scored a set of 100 confirmed ZCT examinations, and three experienced examiners scored a set of 100 MGQT PDD examinations, all of which had resulted from actual criminal investigations. PolyScore 3.3 was used to score the same set of 200 examinations. The examiners scored the examinations using the 7-position scoring scale and those data were later converted to the 3-position scoring scale for comparison. PDD examiner decisions (using both 3- and 7-position scoring) and PolyScore 3.3 decisions were compared to ground truth in order to establish various rates of accuracy for both examiner and algorithm. Interrater agreement for the PDD examiners was calculated, as well as, the proportion of agreement between the individual examiners, the examiners and ground truth, the examiners and PolyScore 3.3, and PolyScore 3.3 and ground truth.

Research Design

This research compared the respective rates of accuracy for PolyScore and two groups of PDD examiners when scoring either a set of 100 ZCT examinations or a set of 100 MGQT examinations (Figure 1).

PDD Examiner Qualifications

Six experienced PDD examiners currently serving as instructors at the DoDPI were designated to blind score the 200 PDD examinations in this study. Due to its bearing on the study methodology, all were familiar with the test data analysis procedures and doctrine currently taught at the DoDPI--to include, response intervals, the 7-position scoring scale, (DoDPI, 1995), and the decision criteria for ZCT and MGQT examinations, (DoDPI, January 1994; DoDPI, November 1995). In addition, each participating PDD examiner had previously performed blind scoring tasks, either as a member of a QC department, or in connection with another research project. None of the examiners had seen the examinations prior to participation in this study, and they were unaware of the total proportions of innocent and guilty cases.

PolyScore 3.3

PolyScore 3.3 was the most recently fielded version of the scoring algorithm at the time this project was completed, and was therefore selected for used in scoring the 200 examinations. The examinations were scored without being manually edited for artifacts, however, PolyScore's own artifact detection system was operating automatically.

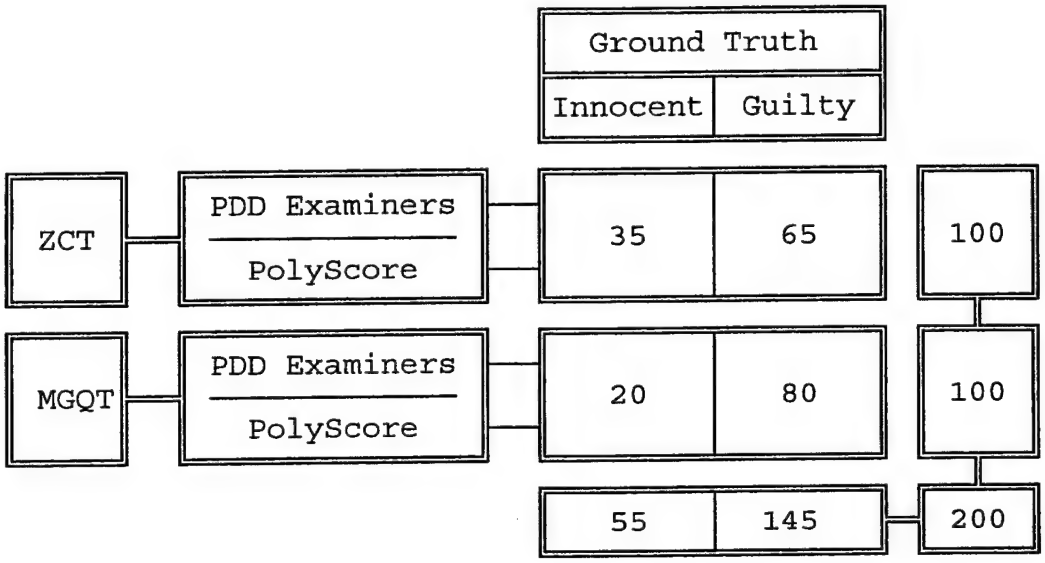


Figure 1. Diagram showing experimental design. *Note.* MGQT = Modified General Question Test; PDD = psychophysiological detection of deception; ZCT = Zone Comparison Test.

PDD Cases

A total of 200 PDD examinations were selected from a database of over 400 confirmed criminal investigations. Due to the small number of confirmed innocent examinations in the database, the study sample was stratified; in that all confirmed innocent exams available were used, and then the remainder of the exams were randomly selected from among the database's confirmed guilty cases. One-hundred (100) examinations utilized the ZCT question format, and the remaining 100 examinations utilized the MGQT question format. The respective examinations were randomly assigned a case number in order to distribute the guilty and innocent cases throughout the data sample.

The database of confirmed cases used in this study is maintained jointly by the APL and the NSA for the purposes of further refining the accuracy of the PolyScore algorithm. The cases themselves were supplied to APL and NSA via an arrangement with seven PDD agencies located within the eastern United States, and are representative of criminal investigations (e.g., murder, arson, larceny, child molestation, drug violations, etc.)

conducted within the agencies' respective areas of operation (Olsen, Harris, Capps, Johnson and Ansley, 1995). All cases were confirmed by, (a) the confession and/or guilty plea of the examinee, or (b) the confession and/or guilty plea of an individual other than the examinee.

Apparatus

A 486 computer, outfitted with the PolyScore software, was used to generate the PolyScore 3.3 decision for each of the 200 examinations. An in-house statistical package known as the Polygraph Research Statistical Package 4.6 (Cestaro, 1995) was later used to perform the various computations associated with the data analyses described in the Results section of this report.

Procedure

Three of the designated PDD examiners were randomly selected to score the 100 ZCT examinations, and the three remaining PDD examiners were tasked with scoring the 100 MGQT examinations. Each participating PDD examiner completed a biographical data form (Appendix A) delineating various aspects of his forensics background. Additionally, all were given a tasking memorandum (Appendix B) which provided the PDD examiners with general project background information, as well as specific scoring instructions.

The cases were selected from the APL/NSA database, and the examination files were reproduced in hard copy. At that time, all participants within each group (ZCT and MGQT) were successively provided with approximately one-third of the printouts to score, until each PDD examiner had scored the complete set of 100 examinations. (Note: Due to the variability in question labeling policies among PDD agencies, the associated question list for each examination was also enclosed in the folder to aid the PDD examiners in accurately determining the comparison and relevant questions.)

Using the provided scoring forms (Appendix C) the PDD examiners from both groups generated numerical evaluations for each relevant question and rendered a decision of DI, INC, or NDI for each examination. In addition, each examiner recorded on the form, the number of minutes spent evaluating the examination (this was done only for internal man-hour accounting purposes). As another administrative exercise separate from the scoring task, the PDD examiners were instructed to notate whether the examination would have been rejected (for reasons of improper tracing size, etc.) had he reviewed it as a QC staff member, rather than as a participant in this research project.

Data Acquisition/Summarization

The scoring/decision criteria for PolyScore 3.3, the 7-position scale, the 3-position scale, and the ZCT and MGQT test formats are detailed below:

PolyScore 3.3. Any examination receiving a probability score of 0.90 or higher was recorded as DI, and any examination receiving a probability score of 0.10 or below was recorded as NDI. All others were labeled as INC examinations. As mentioned previously, no manual editing was performed on the examinations prior to scoring.

7-Position Scoring Scale. A standard seven position scoring scale was used by the examiners when scoring both the ZCT and MGQT examinations. Examiners assigned a value of -3, -2, -1, 0, +1, +2, or +3 to each relevant question, having first compared it to the corresponding comparison question.

3-Position Scoring Scale. The values generated by the examiners using a 7-position scoring scale were later converted to a 3-position scoring scale (-1, 0, +1), by changing the +2 and +3 values to a +1 and the -2 and -3 values to a -1.

ZCT. Figure 2 depicts the question sequence for the ZCT examination, and identifies which questions would be compared during the scoring process. Figure 3 shows the decision criteria and cut off scores for the ZCT examination.

MGQT. Figure 4 depicts the question sequence for the MGQT examination and identifies which questions would be compared during the scoring process. Figure 5 shows the decision criteria and cut off scores for the MGQT examination.

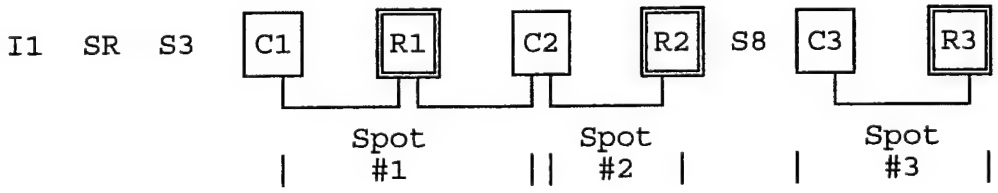


Figure 2. Diagram showing Zone Comparison Test (ZCT) question sequence and evaluation spots with compared comparison and relevant questions linked. *Note.* I = irrelevant; SR = sacrifice relevant; S = symptomatic; C = comparison, and; R = relevant.

Spot#		Score	Call
1 + 2 + 3	≥	+6 (no spot equal to 0 or minus)	NDI
1 + 2 + 3	≤	-6 (no spot equal to 0 or plus)	DI
any	=	-3	DI
1 + 2 + 3	=	any score not mentioned above	INC

Figure 3. Chart showing numerical evaluation criteria for Zone Comparison Test (ZCT) format. *Note.* NDI = no deception indicated; DI = deception indicated; INC = inconclusive.

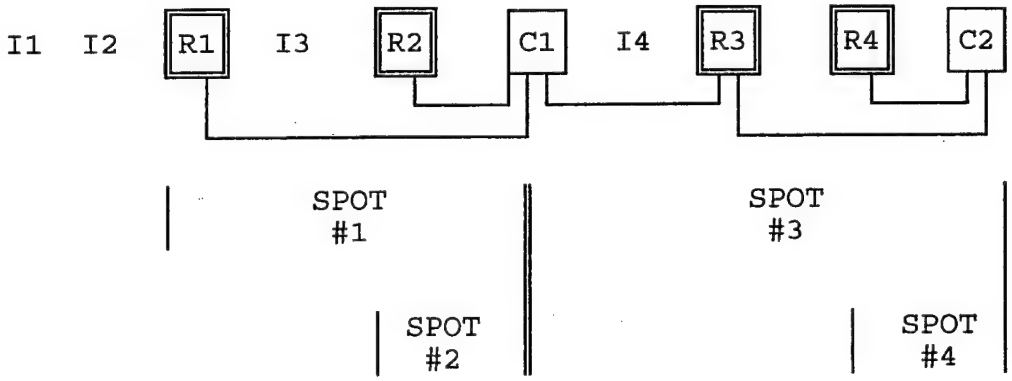


Figure 4. Diagram showing Modified General Question Test (MGQT) question sequence and evaluation spots with compared comparison and relevant questions linked. *Note.* I = irrelevant; C = comparison, and; R = relevant.

Spot#		Score	Call
all	=	+3 or greater	NDI
any	=	-3 or less	DI
1, 2, 3, 4	=	any score not mentioned above	INC

Figure 5. Chart showing numerical evaluation criteria for Modified General Question Test (MGQT) test format. *Note.* NDI = no deception indicated; DI = deception indicated; INC = inconclusive.

Results

PolyScore 3.3 and PDD Examiner Accuracy on ZCT Examinations

Table 1 shows that PolyScore 3.3 was more accurate overall than the PDD examiners as a group, whether they used the 7-position scoring scale or the 3-position scoring scale. PolyScore 3.3 also had a lower percentage of incorrect decisions and a lower INC rate. Using the 7-position scoring scale, the examiners were slightly more accurate (92.3% compared to 85.7%) on the confirmed guilty examinations, but PolyScore 3.3 had more than twice as many correct decisions when scoring the confirmed innocent examinations (93.8% compared to 44.8%). (It should be noted that the calculations involving PolyScore 3.3 are based upon an N = 99, rather than an N = 100, due to the occurrence of a fatal error in the program when attempting to score one guilty examination.)

Due to the higher INC rates generated by the examiners using both the 7- and 3-position scoring scales, their overall accuracy went up when the INC decisions were eliminated from the analysis, however, their error rate went up as well. In most comparisons, the examiners were less accurate and generated both a higher rate of error and a higher rate of INCs when using the 3-position scoring scale. For individual examiner accuracy when using the 7- and 3-position scoring scales see Appendix D.

The results of a test for the significance of proportion differences is shown in Table 2. Regarding the level of overall accuracy, there were a number of comparisons made between PolyScore 3.3 and the examiners where the differences were statistically significant. The same was true, to an even greater extent, when assessing the level of accuracy on the confirmed innocent examinations.

When comparing the differences between the levels of accuracy on the innocent and guilty examinations, only those comparisons involving examiner performance were statistically significant. Table 3 shows that the above statement was true when the examiners used both the 7- and 3-position scales.

Table 1
Percentage of PolyScore 3.3 and PDD Examiner Accuracy on ZCT Examinations When Compared to Ground Truth

			PDD Examiners			
	PolyScore 3.3*		7-Position X		3-Position X	
Decision	%	(n)	%	(n)	%	(n)
With Inconclusives						
Overall (N = 100)						
Correct	90.9	(90)	75.7	(227)	66.3	(199)
Incorrect	3.0	(3)	11.3	(34)	9.0	(27)
Inconclusive	6.1	(6)	13.0	(39)	24.7	(74)
Innocent (n = 35)						
Correct	85.7	(30)	44.8	(47)	31.4	(33)
Incorrect	5.7	(2)	29.5	(31)	23.8	(25)
Inconclusive	8.6	(3)	25.7	(27)	44.8	(47)
Guilty (n = 65)						
Correct	93.8	(60)	92.3	(180)	85.1	(166)
Incorrect	1.6	(1)	1.5	(3)	1.0	(2)
Inconclusive	4.7	(3)	6.2	(12)	13.9	(27)
Without Inconclusives						
Overall						
Correct	96.8	(90)	87.0	(227)	88.1	(199)
Incorrect	3.2	(3)	13.0	(34)	12.0	(27)
Innocent						
Correct	93.8	(30)	60.3	(47)	56.9	(33)
Incorrect	6.3	(2)	39.7	(31)	43.1	(25)
Guilty						
Correct	98.4	(60)	98.4	(180)	98.8	(166)
Incorrect	1.6	(1)	1.6	(3)	1.2	(2)

Note. PDD = psychophysiological detection of deception; ZCT = Zone Comparison Test.
*PolyScore 3.3 accuracy is based on N = 99 due to a fatal error which occurred in the program when attempting to score one guilty examination.

Table 2
Test for the Significance of Proportion Differences When Comparing PolyScore 3.3 and PDD Examiner Decisions (7- and 3-Position Scales) for the ZCT Examinations

	7-Position Scale		3-Position Scale	
Decision	z	p	z	p
Overall				
Correct	2.872	0.004	4.226	0.000
Incorrect	-2.269	0.023	-1.780	0.075
Inconclusive	-1.655	0.098	-3.630	0.000
Innocent				
Correct	4.444	0.000	5.395	0.000
Incorrect	-3.221	0.001	-2.789	0.005
Inconclusive	-2.447	0.014	-3.887	0.000
Guilty				
Correct	-1.199	0.230	0.096	0.923
Incorrect	1.283	0.199	1.487	0.137
Inconclusive	0.521	0.602	-0.952	0.341

Note. PDD = psychophysiological detection of deception; ZCT = Zone Comparison Test.

Table 3
Test for the Significance of Proportion Differences for PolyScore 3.3 and PDD Examiner Decisions When Comparing Innocent and Guilty Decisions for the ZCT Examinations

			PDD Examiners			
	PolyScore 3.3		7-Position		3-Position	
Decision	z	p	z	p	z	p
Correct	1.211	0.226	-5.281	0.000	-5.419	0.000
Incorrect	-0.967	0.334	4.218	0.000	3.804	0.000
Inconclusive	-0.717	0.474	2.763	0.006	3.417	0.001

Note. PDD = psychophysiological detection of deception; ZCT = Zone Comparison Test.

PolyScore 3.3 and PDD Examiner Accuracy on MGQT Examinations

The examiners were more accurate when scoring the MGQT examinations than when scoring the ZCT examinations. Though PolyScore 3.3 was less accurate when scoring the MGQT examinations, Table 4 shows that the level of overall accuracy attained by PolyScore 3.3 was essentially equivalent to that attained by the examiners when using the 7-position scoring scale (79.8% compared to 82.3%). PolyScore was more accurate than the examiners when using the 3-position scoring scale (79.8% compared to 73.3%). (It should be noted that the calculations involving PolyScore are based upon an N = 99, rather than an N = 100, due to the occurrence of a fatal error in the program when attempting to score one guilty examination.)

As with the ZCT examinations, PolyScore 3.3 was slightly less accurate than the examiners when scoring the confirmed guilty examinations (87.3% compared to 96.7%). However, once again PolyScore 3.3 generated twice the number of correct decisions as the examiners when scoring the confirmed innocent examinations (50.0% compared to 25.0%).

When scoring the MGQT examinations, the elimination of the INC decisions had less of an impact on the overall percentage of accuracy than when scoring the ZCT examinations. Examiners using the 7-position scoring scale assigned a decision of INC to 7.0% of the MGQT examinations as compared to 13.0% of the ZCT examinations. The use of the 3-position scoring scale resulted in a 24.7% and 17.7% rate of INCs for the ZCT and the MGQT, respectively. For individual examiner accuracy when using the 7- and 3-position scoring scales see Appendix E.

Table 5 shows that the differences between PolyScore 3.3 and the examiners using the 3-position scoring scale were statistically significant when comparing the percentage of INC decisions generated. When the examiners used the 7-position scoring scale, only the comparisons made involving the percentage of correct and incorrect decisions for the confirmed guilty examinations were statistically significant.

All comparisons shown in Table 6 were statistically significant. There was again, as with the ZCT examinations, a difference in the examiners' handling of the confirmed innocent and confirmed guilty MGQT examinations. In this case, there was also a difference in the scoring performance of PolyScore 3.3.

Table 4
Percentage of PolyScore 3.3 and PDD Examiner Accuracy on MGQT Examinations When Compared to Ground Truth

			PDD Examiners			
	PolyScore 3.3*		7-Position X		3-Position X	
Decision	%	(n)	%	(n)	%	(n)
With Inconclusives						
Overall (N = 100)						
Correct	79.8	(79)	82.3	(247)	73.3	(220)
Incorrect	12.1	(12)	10.7	(32)	9.0	(27)
Inconclusive	8.1	(8)	7.0	(21)	17.7	(53)
Innocent (n = 20)						
Correct	50.0	(10)	25.0	(15)	11.7	(7)
Incorrect	30.0	(6)	53.3	(32)	45.0	(27)
Inconclusive	20.0	(4)	21.7	(13)	43.3	(26)
Guilty (n = 80)						
Correct	87.3	(69)	96.7	(232)	88.8	(213)
Incorrect	7.6	(6)	0.0	(0)	0.0	(0)
Inconclusive	5.1	(4)	3.3	(8)	11.3	(27)
Without Inconclusives						
Overall						
Correct	86.8	(79)	88.5	(247)	89.1	(220)
Incorrect	13.2	(12)	11.5	(32)	10.9	(27)
Innocent						
Correct	62.5	(10)	31.9	(15)	20.6	(7)
Incorrect	37.5	(6)	68.1	(32)	79.4	(27)
Guilty						
Correct	92.0	(69)	96.7	(232)	100.0	(213)
Incorrect	8.0	(6)	3.3	(0)	0.0	(0)

Note. MGQT = Modified General Question Test;
PDD = psychophysiological detection of deception.
*PolyScore 3.3 accuracy is based on N = 99 due to a fatal error which occurred in the program when attempting to score one guilty examination.

Table 5
Test for the Significance of Proportion Differences When Comparing PolyScore 3.3 and PDD Examiner Decisions (7- and 3-Position Scales) for the MGQT Examinations

	7-Position Scale		3-Position Scale	
Decision	z	p	z	p
Overall				
Correct	-0.450	0.653	1.082	0.279
Incorrect	0.311	0.756	0.712	0.476
Inconclusive	0.294	0.769	-2.018	0.044
Innocent				
Correct	1.633	0.102	2.622	0.009
Incorrect	-1.495	0.135	-0.980	0.327
Inconclusive	-0.132	0.895	-1.584	0.113
Guilty				
Correct	-2.188	0.029	-0.292	0.771
Incorrect	2.514	0.012	2.514	0.012
Inconclusive	0.566	0.571	-1.423	0.155

Note. MGQT = Modified General Question Test; PDD = psychophysiological detection of deception.

Table 6
Test for the Significance of Proportion Differences for PolyScore 3.3 and PDD Examiner Decisions When Comparing Innocent and Guilty Decisions for the MGQT Examinations

			PDD Examiners			
	PolyScore 3.3		7-Position		3-Position	
Decision	z	p	z	p	z	p
Correct	-3.709	0.000	-7.524	0.000	-6.978	0.000
Incorrect	2.741	0.006	6.909	0.000	6.290	0.000
Inconclusive	2.180	0.029	2.888	0.004	3.354	0.001

Note. MGQT = Modified General Question Test; PDD = psychophysiological detection of deception.

Interrater Agreement

To assess the extent of interrater agreement, independent of the level of accuracy in relation to ground truth, Kappa (Fleiss [1981] for multiple raters) was computed for the three examiners who scored the ZCT examinations and the three who scored the MGQT examinations. The results for the two formats are shown in Tables 7 and 8, respectively. Interrater statistics are provided for both the 7- and 3-positions scoring scales. The Kappa values indicate the examiners had a moderate to high level of agreement and that the differences in agreement among the examiners was not due to chance.

Table 7
Kappa Statistics for Examiner Agreement When Blind Scoring ZCT Examinations

Scoring Scale	Kappa	SE	z	p
7-Position	0.57	0.044	12.99	0.000
3-Position	0.36	0.044	8.07	0.000

Note. ZCT = Zone Comparison Test.

Table 8
Kappa Statistics for Examiner Agreement When Blind Scoring MGQT Examinations

Scoring Scale	Kappa	SE	z	p
7-Position	0.57	0.045	12.67	0.000
3-Position	0.49	0.052	9.32	0.000

Note. MGQT = Modified General Question Test.

Proportion of Agreement

For comparison, the proportion of concurrence between pairs of evaluators was assessed. Tables 9 and 10 show the proportion of concurrence for the set of ZCT examinations when the examiners used the 7- and 3-position scoring scales. The differences between the 7- and 3-position scoring scales is evidenced here, as it was in the previously reported section on accuracy. Earlier, it was reported that the use of the 3-position scoring scale generated lower accuracy when compared to ground truth. Here, it can be seen that the examiners agreed less among themselves as well, when using the 3-position scoring scale. The proportion of concurrence among examiners when using the 7-position scoring scale ranged from 76.8% to 81.0%. When using the 3-position scale, the range was 58.6% to 70.7%.

Table 9
Proportion of Concurrence Between Pairs of Evaluators for the ZCT - 7-Position Scoring Scale

	Evaluator				
Evaluator	Z-1	Z-2	Z-3	PolyScore 3.3	Ground Truth
Z-1	--	81.0	79.0	76.8	78.0
Z-2	--	--	80.0	71.7	72.0
Z-3	--	--	--	76.8	77.0
PolyScore	--	--	--	--	90.9
Ground Truth	--	--	--	--	--

Note. Z-1 = ZCT Examiner One; Z-2 = ZCT Examiner Two; Z-3 = ZCT Examiner Three; ZCT = Zone Comparison Test.

Table 10
Proportion of Concurrence Between Pairs of Evaluators for the ZCT - 3-Position Scoring Scale

	Evaluator				
Evaluator	Z-1	Z-2	Z-3	PolyScore 3.3	Ground Truth
Z-1	--	67.0	68.0	68.7	70.0
Z-2	--	--	63.0	70.7	70.0
Z-3	--	--	--	58.6	59.0
PolyScore	--	--	--	--	90.9
Ground Truth	--	--	--	--	--

Note. Z-1 = ZCT Examiner One; Z-2 = ZCT Examiner Two; Z-3 = ZCT Examiner Three; ZCT = Zone Comparison Test.

Tables 11 and 12 show the same comparisons as made for the set of MGQT examinations. There is still an obvious difference between the proportion of concurrence generated when the examiners used the 7-position scoring scale as compared to the 3-position scoring scale. The range was 78.8% to 92.0%, and 70.7% to 86.0%, respectively. Interestingly, the values generated by the use of the 3-position scoring scale for the MGQT examinations very nearly equal or exceed those of the ZCT examinations, when using the 7-position scoring system.

Table 11
Proportion of Concurrence Between Pairs of Evaluators for the MGQT - 7-Position Scoring Scale

	Evaluator				
Evaluator	M-1	M-2	M-3	PolyScore 3.3	Ground Truth
M-1	--	88.0	92.0	80.8	84.0
M-2	--	--	92.0	78.8	80.0
M-3	--	--	--	81.8	83.0
PolyScore	--	--	--	--	79.8
Ground Truth	--	--	--	--	--

Note. M-1 = MGQT Examiner One; M-2 = MGQT Examiner 2; M-3 = MGQT Examiner 3; MGQT = Modified General Question Test.

Table 12
Proportion of Concurrence Between Pairs of Evaluators for the MGQT - 3-Position Scoring Scale

	Evaluator				
Evaluator	M-1	M-2	M-3	PolyScore 3.3	Ground Truth
M-1	--	80.0	84.0	71.7	73.0
M-2	--	--	86.0	73.7	75.0
M-3	--	--	--	70.7	72.0
PolyScore	--	--	--	--	79.8
Ground Truth	--	--	--	--	--

Note. M-1 = MGQT Examiner One; M-2 = MGQT Examiner 2; M-3 = MGQT Examiner 3; MGQT = Modified General Question Test.

Discussion

As stated in the Introduction section, the primary intent for conducting this study was to quantify the level of benefit --from an accuracy standpoint--which could be gained by utilizing an automated scoring system such as PolyScore 3.3. The findings reported in the Results section show that the level of benefit could be substantial when scoring ZCT examinations. PolyScore's overall accuracy (scoring both confirmed innocent and confirmed guilty cases) was 90.9%, whereas the PDD examiners mean level of accuracy was 75.7% and 66.3% for the 7- and 3-position scoring scales, respectively. Those differences were statistically significant.

The finding regarding overall level of accuracy contradicts an earlier study which used mock crime (laboratory conducted) examinations, as opposed to the "live" data from actual criminal investigations used in this study. Blackwell (1994) showed PolyScore (known then as PASS 2.0) to be less accurate than the examiners. Also reported in that study was the observation that PolyScore tended to be more accurate when scoring the innocent examinations, and the examiners were more accurate when scoring the guilty examinations. In the current study that trend remained true for the examiners, when scoring either the ZCT or the MGQT examinations, and also for PolyScore 3.3 when scoring the ZCT examinations. However, the computerized algorithm was more accurate when scoring the guilty, rather than the innocent MGQT examinations (87.3% compared to 50.0%).

In a subsequent report, Blackwell (1995) showed that the three versions (2.3, 2.9 and 3.0) of PolyScore issued subsequent to the release of PASS 2.0 were each more accurate than its immediate predecessor. Even so, using the 119 ZCT mock crime examinations, the highest overall level of accuracy generated by the PolyScore algorithm was only 72.3% (using Version 3.0 and unedited data with INCs included). The developers of PolyScore 3.3 have long contended that there is a difference between the physiological reactions generated on mock crime examinations when compared to "live" examinations. The results from the current study seem to support that contention.

As with the other PolyScore-related research conducted by the DoDPI, the data in this investigation have been computed both with and without the INC decisions. In PDD field reporting, the INC decisions are termed "no decision", rather than treated as an error. When the INC decisions were eliminated in this study there was less of a difference between PolyScore 3.3 performance and the performance of the examiners. Unlike previous studies (Blackwell, 1994; Blackwell, 1995), however, this was a result of the examiners producing a higher number of INC decisions which, when eliminated, caused their overall level of accuracy to increase. In past research (Blackwell, 1995), PolyScore's percentage of INC decisions was always higher than the examiners, and had ranged from 11.7% to 20.17%, depending upon which version of the algorithm was used. The PolyScore 3.3 INC rate in this study was 6.1% and 8.1% for the ZCT and the MGQT examinations, respectively.

PolyScore 3.3 generated a higher overall percentage of correct decisions, a lower percentage of incorrect decisions and a lower rate of INC decisions for the ZCT examinations. The percentage of correct decisions generated for the confirmed innocent examinations was more than twice that of the examiners (93.8% compared to 44.8%), and though PolyScore's accuracy on the guilty examinations was below that of the examiners (85.7%

compared to 92.3%) the difference was not statistically significant. In addition, the differences between PolyScore's handling of the innocent and guilty examinations was not statistically significant, unlike the values computed for the examiners when using either the 7- or the 3-position scoring scale. Though accuracy did not reach the levels routinely presented by the APL, from these data it can be concluded that PolyScore 3.3 was more accurate in correctly identifying the ZCT examinations than were the PDD examiners.

When considering PolyScore's performance on the MGQT examinations, the algorithm's level of accuracy essentially equaled that of the examiners. The only differences which were statistically significant were for the number of correct and incorrect decisions on the confirmed guilty examinations (PolyScore's 87.3% and 7.6% compared to the examiners' 96.7% and 0.0%, respectively). In addition, both PolyScore 3.3 and the examiners generated differences on the innocent and guilty cases which were statistically significant.

Without exception, the overall level of accuracy generated by the examiners when using the 7-position scoring scale was higher than when using the 3-position scoring scale. The same was true when looking at the overall percentages for either the innocent examinations or the guilty examinations. The various rates of incorrect and INC decisions varied somewhat, with the 7-position scoring scale generating a higher error rate and the 3-position scoring scale generating a higher INC rate on both the ZCT and the MGQT examinations.

Despite the lower than expected levels of accuracy for the examiners, the extent of interrater agreement was shown to be quite high. Thus, it can be interpreted that the examiners were using essentially the same scoring criteria, however those criteria do not appear to be as effective in distinguishing between the physiological responses generated by guilty and innocent individuals, as the criteria currently being used by PolyScore 3.3--particularly for the ZCT examinations.

The data presented in the proportion of concurrence tables support the observation that the examiners were in high agreement on decisions for both the ZCT and the MGQT examinations when using the 7-position scoring scale, and to a lesser extent when using the 3-position scoring scale. Overall, the individual examiners agreed less often with PolyScore 3.3 than with each other.

Regarding the underlying factors which contributed to the results presented in this report, only a few can be discussed with any confidence. It came as no surprise that PolyScore 3.3 was more accurate when scoring ZCT examinations than when scoring MGQT examinations. The ZCT algorithm was the first developed by

the APL, and preceded the release of the MGQT algorithm by more than five years. One could argue that an algorithm designed to compare two types of questions--relevant and comparison--should be impervious to the number of relevant and comparison questions being scored, and to the test format (the presentation order of the questions) being used. However, just as there exists a specific dynamic in the selection and use of a single-issue versus multiple-issue test format, so too, there exists a specific dynamic required for scoring examinations using those formats. The ZCT algorithm has been refined to a greater extent than the MGQT algorithm and therefore, the MGQT algorithm is not yet as accurate as the ZCT.

The lower than expected examiner accuracy rates can perhaps be attributed to a much discussed phenomenon regarding the activity of blind scoring examinations. Anecdotally, many examiners feel that blind scoring examinations is much more difficult, and therefore much less accurate, than scoring an examination conducted personally. Blackwell (1994) noted that the conducting examiners were 8% to 10% more accurate in their decisions than similarly experienced examiners blind scoring those same examinations without benefit of knowledge of the case facts, or the respective examinee's demographic information. An additional 8% to 10% added to the accuracy levels of the blind scorers in this study would more closely approximate the level of accuracy attained by PolyScore 3.3, and the differences between the two would no longer be statistically significant.

The examiners lodged complaints regarding the quality of the charts themselves. Many of the examinations were deemed to be of poor quality due to inappropriate tracing size, low electrodermal activity, inadequate cardiograph cuff inflation, etc. Two MGQT and five ZCT examinations were rescored after having been adjusted to remove the tracings from the pen stops. Whether PolyScore 3.3 is able to filter such extraneous problems, and the examiners are not, is unclear. The fact remains, however, that despite the clear evidence of poor operations having been used by many of the conducting field examiners, PolyScore 3.3 was able to more accurately score the tracings on those examinations. In addition to the poor quality of the tracings, PolyScore 3.3 had to contend with numerous question intervals which were truncated, and thus did not provide the complete information required for the algorithm to generate an accurate decision.

With regard to the higher level of accuracy afforded by the 7-position scoring scale, this too, was no surprise---for at least two reasons. First, a scale with more assignable values will logically promote greater separation between the items, in this case questions, being evaluated. Critics of the 7-position scale argue that it is too subjective, because there is no defined criteria for which reaction warrants a +2 versus a +3 (as well as a -2 and a -3). In effect, by using the 3-position scale

the subjectivity regarding reaction to the stimulus is removed; the response is either larger, smaller, or the same. Second, the accepted practice throughout the polygraph community is to use the same decision criteria for both the 7- and the 3-position scales. Recently completed research has shown that this may be inappropriate. Krapohl, (in press) found when scoring ZCT examinations, that a decision criteria of +4 and -4 used with a 3-position scale most closely approximated the decision results obtained when using the +6 and -6 criteria with a 7-position scale. In the present study, use of the 7-position scoring scale was clearly of greater benefit than the 3-postion scale, despite the associated increase in subjectivity which came as a result of its use. However, the use of the same scoring criteria for both scoring scales, as is done in the polygraph community, may have unfairly handicapped the 3-position scoring scale.

In summary, PolyScore 3.3 was more effective when scoring the ZCT examinations than the examiners, and it essentially equaled the level of accuracy attained by the examiners when scoring the MGQT examinations. As such, it may be considered an effective tool for use in the polygraph arena, despite the fact that its own level of accuracy has room for improvement---particularly on the guilty examinations. Not assessed in this study was PolyScore's level of accuracy when compared to the conducting examiner. As evidenced in previously cited studies, the conducting examiner tends to be more accurate than similarly qualified examiners when generating decisions on a polygraph examination. If that is consistently true, PolyScore's comparative level of accuracy may be reduced except in a QC setting.

The examiners were more accurate when scoring the MGQT examinations than when scoring the ZCT examinations. Without further research, however, it would be hasty to endorse the use of one format over the other. With regard to the 7-position scoring scale, the initial indication is that it is a more accurate method of scoring polygraph examinations than the 3-position scoring scale. There are also complications with endorsing it as the preferred method, due to the prevailing use of the "minus three (-3) in a spot equals a DI decision" scoring criteria. Many of the DI decisions made in this study were the result of a minus three (-3) on a single question. It is unknown whether the 7-position scoring scale would still be more accurate than the 3-position scoring scale, were the "-3" scoring criteria eliminated. Currently proposed research should settle the 7- versus 3-issue.

References

Ansley, N., Garwood, M. & Barland, G. H. (1984). The accuracy and utility of polygraph testing. Washington DC: Department of Defense. (Reprinted in Polygraph, 4, 1-143).

Barland, G. H. (1972). Reliability of polygraph chart evaluations. Polygraph, 1(4), 192-206.

Ben-Shakhar, G., Lieblich, I. & Bar-Hillel, M. (1982). An evaluation of polygraphers' judgments: A review from a decision theoretic perspective. Journal of Applied Psychology, 67(6), 701-713.

Blackwell, N. J. (1994). An evaluation of the effectiveness of the polygraph automated scoring system (PASS) in detecting deception in a mock crime analog study. (Report No. DoDPI94-R-0003). Fort McClellan, AL: Department of Defense Polygraph Institute.

Blackwell, N. J. (1995). POLYSCORE: A comparison of accuracy. (Report No. DoDPI95-R-0001). Fort McClellan, AL: Department of Defense Polygraph Institute.

Capps, M. (1993, January). Polygraph automated scoring system (PASS) version 2.0. Paper presented at the Department of Defense Polygraph Institute training program on PASS Operations, Fort McClellan, AL.

Cestaro, V. (1995). Polygraph research statistical package - Version 4.6 [Computer program]. Fort McClellan, AL.

Department of Defense Polygraph Institute. (1995). Test data analysis. (Available from [Department of Defense Polygraph Institute, Building 3195, Fort McClellan, AL]).

Department of Defense Polygraph Institute. (1994, January). FSC 502 - Zone comparison test (ZCT). (Available from [Department of Defense Polygraph Institute, Building 3195, Fort McClellan, AL]).

Department of Defense Polygraph Institute. (1995, November). FSC 502 - Modified general question technique (MGQT). (Available from [Department of Defense Polygraph Institute, Building 3195, Fort McClellan, AL]).

Fleiss, J. L. (1981). Statistical methods for rates and proportions. New York: John Wiley & Sons.

Forman, R. F. & McCauley, C. (1986). Validity of the positive control polygraph test using the field practice model. Journal of Applied Psychology, 71(4), 691-698.

Furedy, J. J. (1985). In my opinion....Credulous vs. critical police use of the polygraph in criminal investigations. Canadian Journal of Criminology, 27(4), 491-495.

Furedy, J. J. (1987). Evaluating polygraphy from a psychophysiological perspective - A specific effects analysis. Pavlovian Journal of Biological Science, 22(4), 145-152.

Iacono, W. G. (1991). Can we determine the accuracy of polygraph tests? In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), Advances in psychophysiology (Vol. 4). London: Jessica Kingsley Publishers.

The Johns Hopkins University Applied Physics Laboratory. (1996). POLYSCORE--Version 3.3 [Computer program]. Laurel, MD.

Krapohl, D. J. (in press). A comparison of 3- and 7-position scoring scales with laboratory data. Polygraph.

Kircher, J. C. & Raskin, D. C. (1988). Human versus computerized evaluations of polygraph data in a laboratory setting. Journal of Applied Psychology, 73(2), 291-302.

Lykken, D. T. (1981). A tremor in the blood: Uses and abuses of the lie detector. New York: McGraw-Hill.

Lykken, D. T. (1986, April) Polygraph testing: "20th-century witchcraft"? [Letter to the editor] Medical Economics, p. 20.

Lykken, D. T. (1988). The case against polygraph testing. In A. Gale (Ed.), The polygraph test: Lies, truth and science (pp. 111-125). London: Sage Publications, Inc.

Lykken, D. T. (1991). The lie detector controversy: An alternate solution. In J. R. Jennings, P. K. Ackles, & M. G. H. Coles (Eds.), Advances in psychophysiology (Vol. 4, pp. 209-214). London: Jessica Kingsley Publishers.

Office of Technology Assessment. (1983). Scientific validity of polygraph testing: A research review and evaluation-A technical memorandum (Report No. OTA-TM-H-15). Washington, DC: U.S. Government Printing Office.

Olsen, D. E., Harris, J. C., Capps, M. H., Johnson, G. J. & Ansley, N. (1995). Computerized polygraph scoring system. Unpublished manuscript.

Raskin, D. C. (1979, December). Science and the polygraph profession. Texas Association of Polygraph Examiners Newsletter, pp. 3-7.

Raskin, D. C. (1988). Does science support polygraph testing? In A. Gale (Ed.), The polygraph test: Lies, truth and science, (pp. 96-110). London: Sage Publications.

Appendix A

BLIND SCORER BIOGRAPHICAL DATA	
Name: _____	Age: _____
Current employer/agency: _____	
Current job series/title: _____	
PDD school attended for initial certification training: Name _____ Location _____	
Date of PDD certification (mm/yy): _____/_____	
Certifying authority/agency: _____	
Total # PDD exams conducted as of 31 DEC 1995: Criminal _____ (approximate) Counterintelligence _____ (approximate)	
Case resolution rate as a field examiner: _____%	
Have you completed the DoDPI Axciton course? No _____ Yes _____ Date_____/_____ mm / yy	

Have you conducted "live" exams using an Axciton?

No _____ Yes _____ # _____ (approximate)

Have you blind scored Axciton exams prior to this study?

No _____ Yes _____ # _____ (approximate)

Length of time as:

	Years	Months
DoDPI instructor	_____	_____
PDD examiner		
Criminal	_____	_____
Counterintelligence	_____	_____
PDD quality control		
Criminal	_____	_____
Counterintelligence	_____	_____
Investigator		
Criminal	_____	_____
Counterintelligence	_____	_____

Appendix B

Blind Scorer Tasking Memorandum

MEMORANDUM FOR FACULTY

SUBJECT: Blind scoring of psychophysiological detection of deception (PDD) examinations in support of DoDPI research project # DoDPI96-P-0001.

1. You have been designated to blind score a set of 100 confirmed PDD examinations in support of a project entitled, "POLYSCORE and Psychophysiological Detection of Deception (PDD) Examiner Rates of Accuracy When Scoring Examinations from Actual Criminal Investigations", (DoDPI96-P-0001). Your respective rate of accuracy will be compared with the level of accuracy attained by the POLYSCORE algorithm when scoring the same set of examinations. NOTE: Accuracy rates for both POLYSCORE, and the individual examiners will be posted on the small bulletin board in the hallway (near the lounge) when all examinations have been scored.

2. Three examiners were randomly selected to score Modified General Question Technique (MGQT) examinations and an additional three examiners were selected to score Zone Comparison Technique (ZCT) examinations (see distribution list below).

3. The examinations were supplied by a number of different agencies and will, therefore, vary in regard to format version (e.g. ZCT and Bi-zone), question labeling procedures, and overall tracing quality. All printouts were generated at DoDPI, using the sensitivity settings selected by the original examiner. Due to unclear question labeling on some cases, the question list for each examination has been included in the folder to aid you in identifying controls and relevants.

4. Along with this memo, you have been provided with a packet of scoring forms and a personal log sheet to track your decision for each completed examination, if desired. At your earliest convenience, please stop by my office (E 108), to pick up your first set of examinations. You will initially be given one-third of the examinations, which you may score and return to me or, alternately, pass on to another examiner on the same distribution list as yourself. However, please return your scoring forms directly to me.

5. Scoring criteria and response windows used during this project should comply with DoDPI doctrine, as taught in the Basic Courses in Forensic Psychophysiology program. If you have questions regarding any aspect of those published guidelines, please see me.

6. Specific scoring instructions:

- o Do not mark on the examinations.
- o Insure that both the folder and the printout inside the folder are labeled with the same number (e.g., M-001, Z-001, etc.) and record that number as the Case# at the top of the scoring form.
- o Use DoDPI scoring criteria and response windows.
- o Use 7-position scoring scale.
- o Provide a decision (DI, INC, or NDI) for each examination,
  - however, if you feel that the examination was of inferior quality and that, given the opportunity, you would have rejected it, please indicate that at the top of the scoring form by writing the word "REJECT".
- o In the top right corner of the scoring form, write the estimated number of minutes required to score each respective examination (this is required for project manhour accounting purposes).
- o Return your scoring forms directly to me, NLT 15 JAN 1996.

7. If you have other questions about this project, please contact me at x6894, or stop by my office (E 108).

ORIGINAL SIGNED

N. JOAN BLACKWELL
Research Psychologist

Appendix C

Scoring Form

Identification # _____ Case # _____

Component	Q#	Q#	Q#	Q#
Pneumo1				
Pneumo2				
GSR				
Cardio				
Chart Sub-Totals				

Component	Q#	Q#	Q#	Q#
Pneumo1				
Pneumo2				
GSR				
Cardio				
Chart Sub-Totals				

Component	Q#	Q#	Q#	Q#
Pneumo1				
Pneumo2				
GSR				
Cardio				
Chart Sub-Totals				

Grand Total _____ (ZCT Only)

Decision _____

(Note: Form reduced to fit on page.)

Appendix D

Individual Examiner Accuracy on the ZCT Examinations

Table D-1
Percentage of PDD Examiner Agreement with Ground Truth on ZCT Examinations - 7-Position Scoring Scale

	PDD Examiner					
	Z-1		Z-2		Z-3	
Decision	%	(n)	%	(n)	%	(n)
With Inconclusives						
Overall (N = 100)						
Correct	78.0	(78)	72.0	(72)	77.0	(77)
Incorrect	10.0	(10)	14.0	(14)	10.0	(10)
Inconclusive	12.0	(12)	14.0	(14)	13.0	(13)
Innocent (n = 35)						
Correct	54.3	(19)	28.6	(10)	51.4	(18)
Incorrect	25.7	(9)	40.0	(14)	22.9	(8)
Inconclusive	20.0	(7)	31.4	(11)	25.7	(9)
Guilty (n = 65)						
Correct	90.8	(59)	95.4	(62)	90.8	(59)
Incorrect	1.5	(1)	0.0	(0)	3.1	(2)
Inconclusive	7.7	(5)	4.6	(3)	6.2	(4)
Without Inconclusives						
Overall						
Correct	88.6	(78)	83.7	(72)	88.5	(77)
Incorrect	11.4	(10)	16.3	(14)	11.5	(10)
Innocent						
Correct	67.9	(19)	41.7	(10)	69.2	(18)
Incorrect	32.1	(9)	58.3	(14)	30.8	(8)
Guilty						
Correct	98.3	(59)	100.0	(62)	96.7	(59)
Incorrect	1.7	(1)	0.0	(0)	3.3	(2)

Note. PDD = psychophysiological detection of deception; Z-1 = ZCT Examiner One; Z-2 = ZCT Examiner Two; Z-3 = ZCT Examiner Three; ZCT = Zone Comparison Test.

Table D-2
Percentage of PDD Examiner Agreement with Ground Truth on ZCT Examinations - 3-Position Scoring Scale

	PDD Examiner					
	Z-1		Z-2		Z-3	
Decision	%	(n)	%	(n)	%	(n)
With Inconclusives						
Overall (N = 100)						
Correct	70.0	(70)	70.0	(70)	59.0	(59)
Incorrect	8.0	(8)	12.0	(12)	7.0	(7)
Inconclusive	22.0	(22)	18.0	(18)	34.0	(34)
Innocent (n = 35)						
Correct	42.9	(15)	22.9	(8)	28.6	(10)
Incorrect	20.0	(7)	34.3	(12)	17.1	(6)
Inconclusive	37.1	(13)	42.9	(15)	54.3	(19)
Guilty (n = 65)						
Correct	84.6	(55)	95.4	(62)	75.4	(49)
Incorrect	1.5	(1)	0.0	(0)	1.5	(1)
Inconclusive	13.9	(9)	4.6	(3)	23.1	(15)
Without Inconclusives						
Overall						
Correct	89.7	(70)	85.4	(70)	89.4	(59)
Incorrect	10.3	(8)	14.6	(12)	10.6	(7)
Innocent						
Correct	68.2	(15)	40.0	(8)	62.5	(10)
Incorrect	31.8	(7)	60.0	(12)	37.5	(6)
Guilty						
Correct	98.2	(55)	100.0	(62)	98.0	(49)
Incorrect	1.8	(1)	0.0	(0)	2.0	(1)

Note. PDD = psychophysiological detection of deception; Z-1 = ZCT Examiner One; Z-2 = ZCT Examiner Two; Z-3 = ZCT Examiner Three; ZCT = Zone Comparison Test.

Appendix E

Individual Examiner Accuracy on the MGQT Examinations

Table E-1
Percentage of PDD Examiner Agreement with Ground Truth on MGQT Examinations - 7-Position Scoring Scale

	PDD Examiner					
	M-1		M-2		M-3	
Decision	%	(n)	%	(n)	%	(n)
With Inconclusives						
Overall (N = 100)						
Correct	84.0	(84)	80.0	(80)	83.0	(83)
Incorrect	9.0	(9)	12.0	(12)	11.0	(11)
Inconclusive	7.0	(7)	8.0	(8)	6.0	(6)
Innocent (n = 20)						
Correct	30.0	(6)	20.0	(4)	25.0	(5)
Incorrect	45.0	(9)	60.0	(12)	55.0	(11)
Inconclusive	25.0	(5)	20.0	(4)	20.0	(4)
Guilty (n = 80)						
Correct	97.5	(78)	95.0	(76)	97.5	(78)
Incorrect	0.0	(0)	0.0	(0)	0.0	(0)
Inconclusive	2.5	(2)	5.0	(4)	2.5	(2)
Without Inconclusives						
Overall						
Correct	90.3	(84)	87.0	(80)	88.3	(83)
Incorrect	9.7	(9)	13.0	(12)	11.7	(11)
Innocent						
Correct	40.0	(6)	25.0	(4)	31.3	(5)
Incorrect	60.0	(9)	75.0	(12)	68.8	(11)
Guilty						
Correct	100.0	(78)	100.0	(76)	100.0	(78)
Incorrect	0.0	(0)	0.0	(0)	0.0	(0)

Note. M-1 = MGQT Examiner One; M-2 = MGQT Examiner 2; M-3 = MGQT Examiner 3; MGQT = Modified General Question Test; PDD = psychophysiological detection of deception.

Table E-2
Percentage of PDD Examiner Agreement with Ground Truth on MGQT Examinations - 3-Position Scoring Scale

	PDD Examiner					
	M-1		M-2		M-3	
Decision	%	(n)	%	(n)	%	(n)
With Inconclusives						
Overall (N = 100)						
Correct	73.0	(73)	75.0	(75)	72.0	(72)
Incorrect	6.0	(6)	10.0	(10)	11.0	(11)
Inconclusive	21.0	(21)	15.0	(15)	17.0	(17)
Innocent (n = 20)						
Correct	10.0	(2)	10.0	(2)	15.0	(3)
Incorrect	30.0	(6)	50.0	(10)	55.0	(11)
Inconclusive	60.0	(12)	40.0	(8)	30.0	(6)
Guilty (n = 80)						
Correct	88.8	(71)	91.3	(73)	86.3	(69)
Incorrect	0.0	(0)	0.0	(0)	0.0	(0)
Inconclusive	11.3	(9)	8.8	(7)	13.8	(11)
Without Inconclusives						
Overall						
Correct	92.4	(73)	88.2	(75)	86.8	(72)
Incorrect	7.6	(6)	11.8	(10)	13.3	(11)
Innocent						
Correct	25.0	(2)	16.7	(2)	21.4	(3)
Incorrect	75.0	(6)	83.3	(10)	78.6	(11)
Guilty						
Correct	100.0	(71)	100.0	(73)	100.0	(69)
Incorrect	0.0	(0)	0.0	(0)	0.0	(0)

Note. M-1 = MGQT Examiner One; M-2 = MGQT Examiner 2; M-3 = MGQT Examiner 3; MGQT = Modified General Question Test; PDD = psychophysiological detection of deception.